Chapter 1
Accelerated Biological Simulation Research (ABSR)

Societal Needs for Accelerated Simulation Research in Biology

The pace of extraordinary advances in molecular biology has accelerated in the past decade due in large part to discoveries coming from genome projects on human and model organisms. The advances in the genome project so far, happening well ahead of schedule and under budget, have exceeded any dreams by its protagonists, let alone formal expectations. Biologists expect the next phase of the genome project to be even more startling in terms of dramatic breakthroughs in our understanding of human biology, the biology of health and of disease.

Why can we not wait for a computational science effort until the conclusion of the genome project, and what are the broader implications? Significant personal and economic costs will be borne by society for delays in our exploiting the discoveries of the genome projects on behalf of the Nation. The importance of each individual, each human being on the planet, the individual's productivity and contribution to society, the quality of that individual's life including that of our shared environment, has transcendental value. A major break through in technology, the human genome project, puts us on the brink of truly extraordinary progress in realizing the transcendental goal of human well-being. Only today can biologists begin to envision the necessary experimental, computational and theoretical steps necessary to exploit genome sequence information for its medical impact, its contribution to biotechnology and economic competitiveness, and its ultimate contribution to environmental quality.

Individualized medicine- the recognition of individual differences in drug and treatment response, in disease development and progression, in the appearance and thus diagnosis of disease- justifies a much more aggressive approach to utilizing DNA sequence information and the introduction of simulation capabilities to extract the implicit information contained in the human genome. Each year thousands die and a hundred times more people suffer adverse reactions of various extents to drugs that are applied in perfectly correct usage, dose and disease specificity. Cancer in particular stands out as a disease of the genes. Two patients with what appear to be identical cancers (based on cellular pathology) at the same stage of malignancy, growth and dissemination will have very different responses to the same therapy regime; for example, one could rapidly fail to respond to treatment, become terminal in weeks, and the other could respond immediately and ultimately recover fully.

Re-engineering microbes for bioremediation will depend directly on the understanding derived from the simulation studies proposed here. The design of new macromolecules and the redesign of microbe metabolism based on simulation studies, truly grand challenges for applied biology, will contribute to a wide range of environmental missions for the Department, including the environmental bioremediation of the Nation's most contaminated sites, often mixed waste for which no economically feasible cleanup technology currently exists. Many aspects of sustainable development, changing the nature of industrial processes to use environmentally friendly processes, depend on the same kinds of advances.

Similarly, in the longer term the understanding of living systems will also contribute to environmental research on carbon management.

To exploit the inherent genome information derived from knowing the DNA sequences, computational advances, along with related experimental biotechnology, are essential. Knowing the sequence of the DNA does not tell us about the function of the genes, specifically the actions of their protein products - where, when, why, how the proteins act is the essence of biological knowledge required. Encoded in the DNA sequence is a protein's three dimensional topography, which in turn determines function; uncovering this sequence-structure-function relationship is the core goal of modern structural biology today. The goal of the accelerated computational biology initiative is to link sequence, structure and function, and to move from analysis of individual macromolecules to macromolecular assemblies and complex oligomeric interactions that make up the complex processes within the cell.

Elucidating the Path From Protein Structure, to Function, to Disease

A particular sequence derived from the genome encodes in a character string the three-dimensional structure of a protein that performs a specific function in the cell. The primary outcome of the Argonne Workshop on Structural Genomics emphasized the goal of exploiting the relationship between sequence and structure at a level unattainable before the coordinated effort to map and sequence the genomes of many organisms. It is in fact a logical extension of the genome effort to systematically elaborate DNA sequences into full three dimensional structures and functional analysis; this new effort will require the same level of cooperation and collaboration among scientists as was necessary in the original genome project.

The Structural Genomics Initiative is gaining momentum and will primarily focus initially on the infrastructure and support necessary to realize high throughput experimental structures by x-ray crystallography and high field NMR. Success of this effort is in some sense guaranteed, and the computational challenges that are posed when a structural genomics initiative of this scale is contemplated is already clear cut. It is the outgrowth from this direction, a computational biotechnology initiative- that will capture everything from sequence, structure and functional genomics to genetic networks and metabolic engineering to forward folding of macromolecules and kinetics to cellular level interactions. This approach makes all the modeling and simulation of macromolecules fair game and moves toward the complex systems approach, the complicated level of real living systems rather than individual macromolecules.

Elucidating the path from structure, to function, to epidemiological consequences for proteins of selected pathogens directly relates structural analysis to national health. As sequence information continues to accumulate at a rapid pace from the Genome Program, and as structure information becomes increasingly available from the Structural Genomics Initiative, the challenge will be to integrate biological information from the molecular level of sequence and structure, to the macroscopic level of function. Such an integrated approach lies at the heart of our ability to understand and combat disease. Since the early days of the Genome Project DOE has been a pioneer in establishing databases of molecular information and in developing algorithms and computational tools to analyze this information. In the

past, this effort has concentrated on the human genome, with the goal of understanding basic molecular mechanisms central to biological function. Establishing such an understanding is critical to developing a molecular interpretation of disease, and augmenting these database with structural information from the Structural Genomics Initiative will prove immensely valuable.

An Accelerated Biological Simulation Research Initiative

The goal of the ABSR effort within the time horizon of the Strategic Simulation Initiative is to:

· ***Characterize the link between protein sequence and fold topology.*** The emphasis on the experimental determination of a complete set of representative tertiary folds is based on the success of comparative modeling. One can often deduce the fold topology for a new sequence by simply finding another similar sequence with a known structure, even when sequence identity is very low. The ~100,000 protein coding genes expected in the human genome is too large to handle experimentally, but protein modelers estimate that ~10,000 structures for protein domains would be sufficient to use these algorithms to determine the fold topology of the remaining 90% of gene products. Algorithms of scale have been developed that attempt to find this sequence-fold match, but more sophisticated versions that exhibit more severe scaling are needed to be genuinely successful in this regard.

· ***Quantitative determination of protein structure from folding or conformational searches.*** Although the gross backbone configuration (or tertiary fold) of proteins remains invariant under sequence mutation, quantitative differences in structure between wild type and mutant can have important macroscopic effects on protein function. The quantitative prediction of structure and folding behavior is critical for the development of successful pharma-ceutical drug targets, protein redesign of enzymes for bioremediation, and prediction of disease manifestation of new pathogens. The next step beyond predicting fold topology is the quantitative determination of protein structure starting first from the tertiary fold prediction, and ultimately directly from sequence. This is the heart of computational complexity in molecular biology at present. The simulation methodologies have commonality with the areas of materials and combustion chemistry, and share the same severe scaling issues due to large length scales, long time scales, and system size scaling that define the need for high-end computing in quantitative modeling.

· ***Simulate the biochemical function of individual gene products.*** A robust and predictive approach to protein structure ties directly into our ability to model and understand protein function. Simulations will be important for predicting detailed structural changes and the fluctuations that drive enzymatic reactions, how protein structures recognize each other, and associate to form the multi-protein complexes, and prediction of the mode and energies with which molecules bind to proteins in metabolic reactions and molecular signaling processes. The strong biological connection between structure and function means that the same modeling issues, primarily a good energy surface description and the means to explore it, will be important for simulating biochemical function as well.